FoML08 Nguyen Ngoc Nhu Y July 07, 2025

FINAL PROJECT REPORT

Table of Contents

A. REGRESSION MODEL 2
1. OVERVIEW 2
2. EXPLORATORY DATA ANALYSIS 2
2.1. PRE-TRANSFORM EDA 2
2.2. POST-TRANSFORM EDA 5
3. PIPELINE DESCRIPTION 8
4. RESULT & EVALUATION 9
B. CLASSIFICATION MODEL 9
5. OVERVIEW 9
6. EXPLORATORY DATA ANALYSIS 10
6.1. PRE-TRANSFORM EDA 10
6.2. POST-TRANSFORM EDA 22
7. PIPELINE DESCRIPTION 26
8. RESULT & EVALUATION 27

A. REGRESSION MODEL

1. OVERVIEW

Model: Predict used car's price.

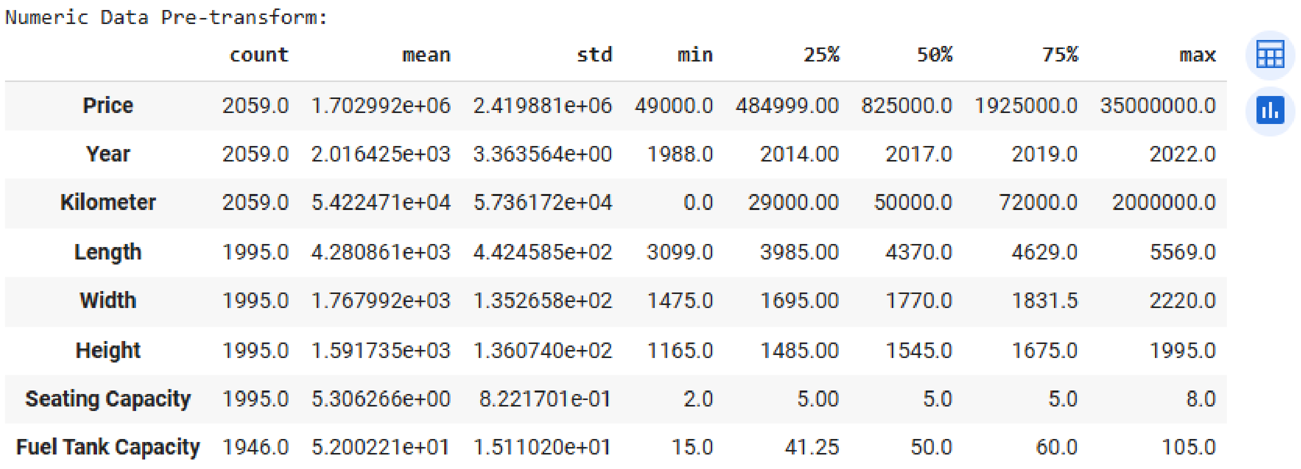
Data origin: car data set (version 4) updated 2 years ago by Nihal Birla, Nishant Verma, Nikhil Kushwaha via Kaggle platform.

2. EXPLORATORY DATA ANALYSIS

2.1. PRE-TRANSFORM EDA

Total 2059 entries with 5 features are float64, 3 features are int64, 12 features are object.

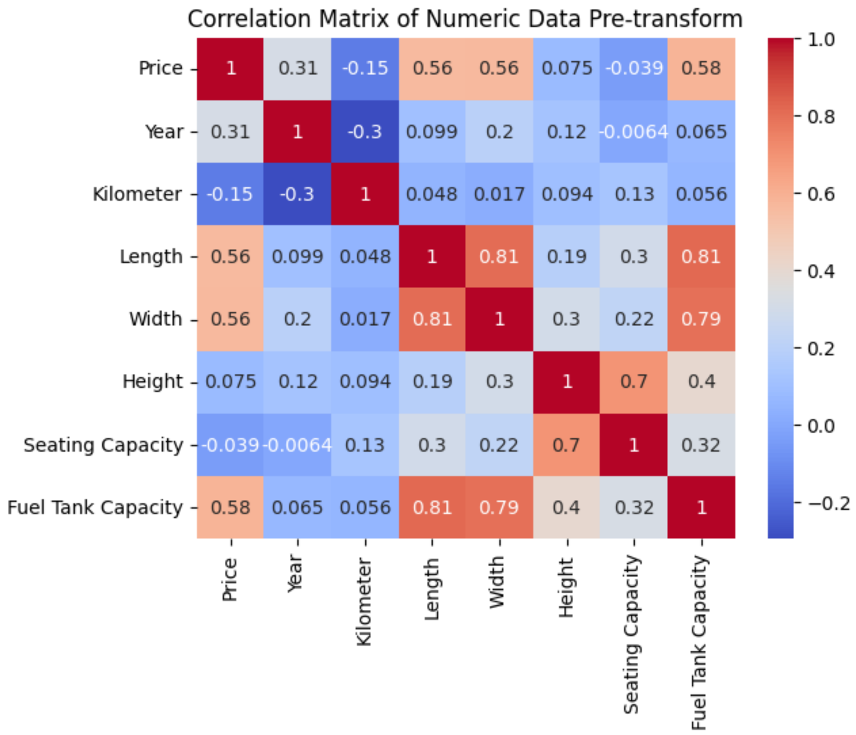
2.1.1. NUMERIC FEATURES



```
Numeric Data Pre-transform:
```

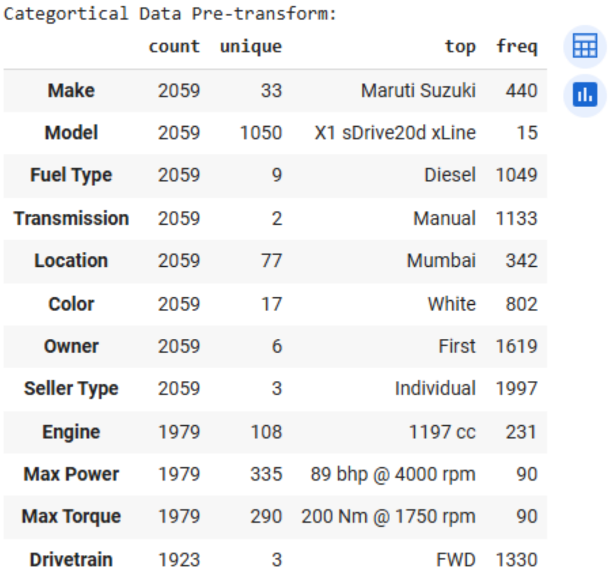
	count	mean	std	min	25%	50%	75%	max
Price	2059.0	1.702992e+06	2.419881e+06	49000.0	484999.00	825000.0	1925000.0	35000000.0
Year	2059.0	2.016425e+03	3.363564e+00	1988.0	2014.00	2017.0	2019.0	2022.0
Kilometer	2059.0	5.422471e+04	5.736172e+04	0.0	29000.00	50000.0	72000.0	2000000.0
Length	1995.0	4.280861e+03	4.424585e+02	3099.0	3985.00	4370.0	4629.0	5569.0
Width	1995.0	1.767992e+03	1.352658e+02	1475.0	1695.00	1770.0	1831.5	2220.0
Height	1995.0	1.591735e+03	1.360740e+02	1165.0	1485.00	1545.0	1675.0	1995.0
Seating Capacity	1995.0	5.306266e+00	8.221701e-01	2.0	5.00	5.0	5.0	8.0
Fuel Tank Capacity	1946.0	5.200221e+01	1.511020e+01	15.0	41.25	50.0	60.0	105.0

- ➢ Price and Max Power/Max Torque/Engine: There are strong positive correlations between Price and Max Power, Max Torque, and Engine. This is expected, as cars with more powerful engines and higher torque generally tend to be more expensive.
- ➢ Price and Length/Width/Fuel Tank Capacity: There are also positive correlations between Price and Length, Width, and Fuel Tank Capacity. Larger cars with bigger fuel tanks might be associated with higher price points, possibly indicating larger or more luxurious vehicle segments.
- ➢ Age and Price/Length/Width: Age shows a negative correlation with Price, Length, and Width. This makes sense, as older cars are typically less expensive and car dimensions have changed over time.
- ➢ Seating Capacity: Seating Capacity does not appear to have a strong correlation with Price or most other numerical features in this initial view.



2.1.2. CATEGORICAL FEATURES

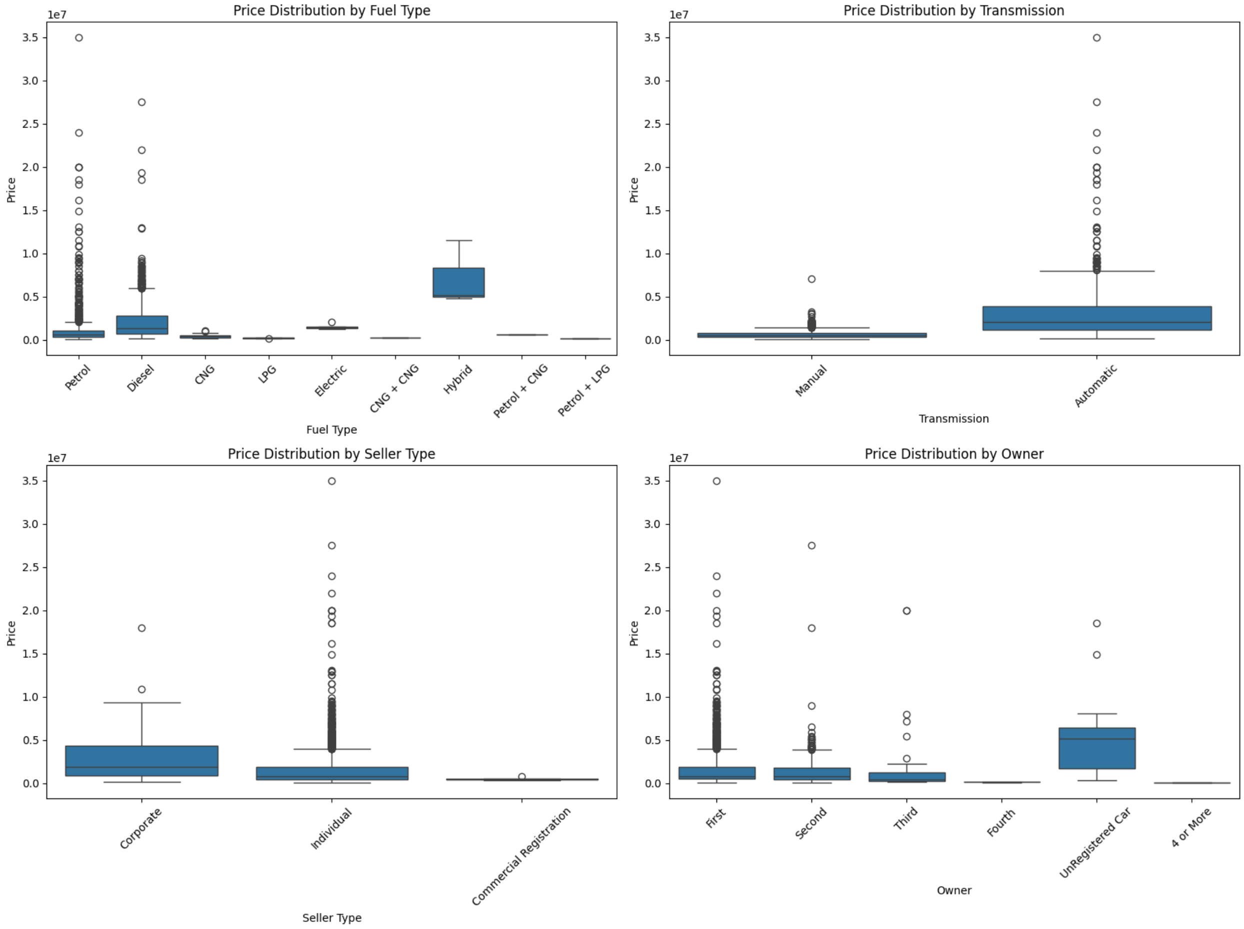
- For feature Model, the number of unique values is nearly half of the total number of samples. This variable can almost be considered an identifier or close to a highly personalized characteristic. When processing this variable, applying the One-Hot Encoding technique can create a large number of new features, leading to a complex model and causing the "curse of dimensionality".



Categortical Data Pre-transform:

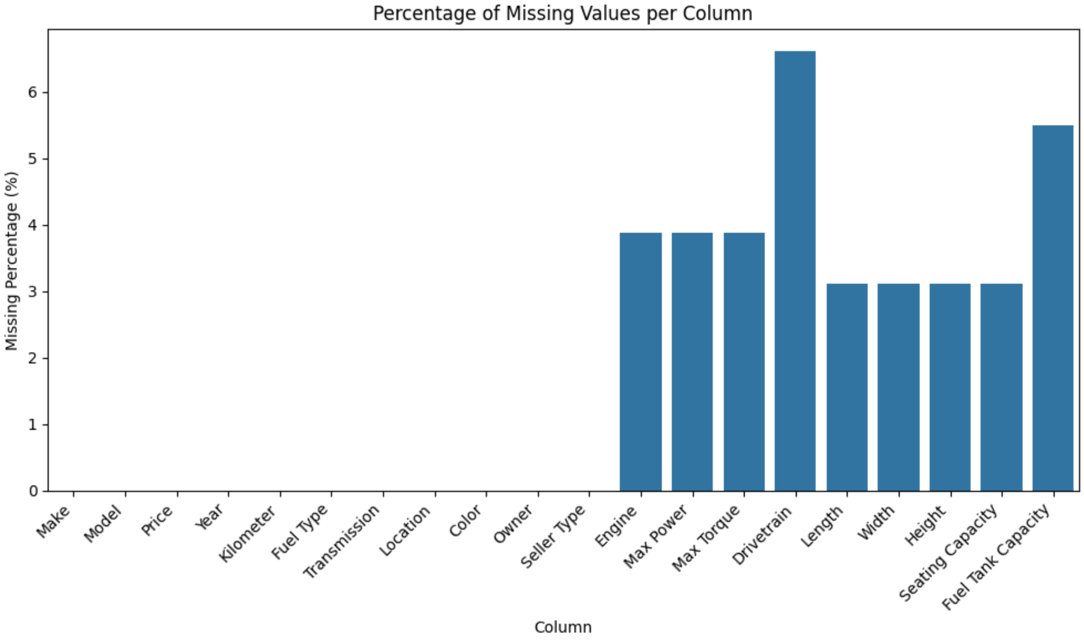
	count	unique	top	freq
Make	2059	33	Maruti Suzuki	440
Model	2059	1050	X1 sDrive20d xLine	15
Fuel Type	2059	9	Diesel	1049
Transmission	2059	2	Manual	1133
Location	2059	77	Mumbai	342
Color	2059	17	White	802
Owner	2059	6	First	1619
Seller Type	2059	3	Individual	1997
Engine	1979	108	1197 cc	231
Max Power	1979	335	89 bhp @ 4000 rpm	90
Max Torque	1979	290	200 Nm @ 1750 rpm	90
Drivetrain	1923	3	FWD	1330

- Fuel Type: The box plot for Fuel Type shows that Diesel and Petrol cars tend to have higher median prices and a wider range of prices compared to other fuel types like CNG or LPG. Electric cars also show a significant range of prices.
- Transmission: Cars with Automatic transmission generally have higher median prices and a larger spread of prices compared to Manual transmission cars. This suggests that automatic cars are often in higher price segments.
- Seller Type: The Individual seller type has a much wider distribution of prices, including many outliers at higher prices, compared to Dealer and Trustmark Dealer. This is expected as individual sellers might have a more varied range of cars and pricing strategies.
- Owner: The box plot for Owner shows that cars with First owners tend to have the highest median price and the widest price range. As the number of owners increases (Second, Third, Fourth & Above), the median price generally decreases, which is intuitive as cars depreciate with more owners.



2.1.3. MISSING VALUE

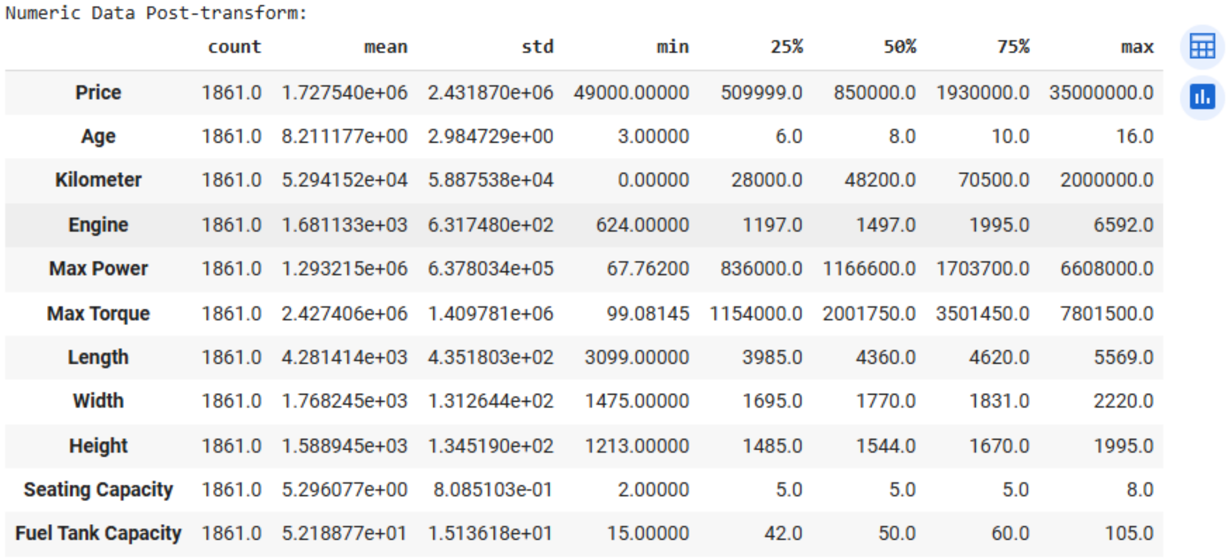
- Drivetrain and Fuel Tank Capacity are the columns with the most significant amount of missing data, followed by the engine-related features and dimensions/seating capacity.



2.2. POST-TRANSFORM EDA

After (1) dropping missing value/outliers/unnecessary features, (2) transform a feature, remains 1874 entries with 5 features are float64, 4 features are int64, 12 features are object.

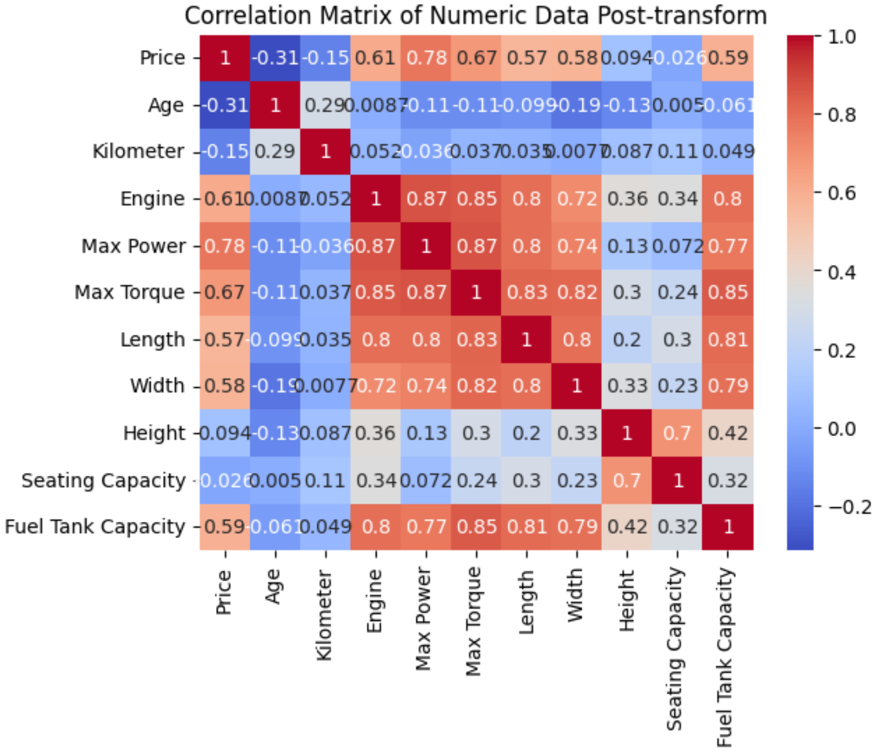
2.2.1. NUMERIC FEATURES



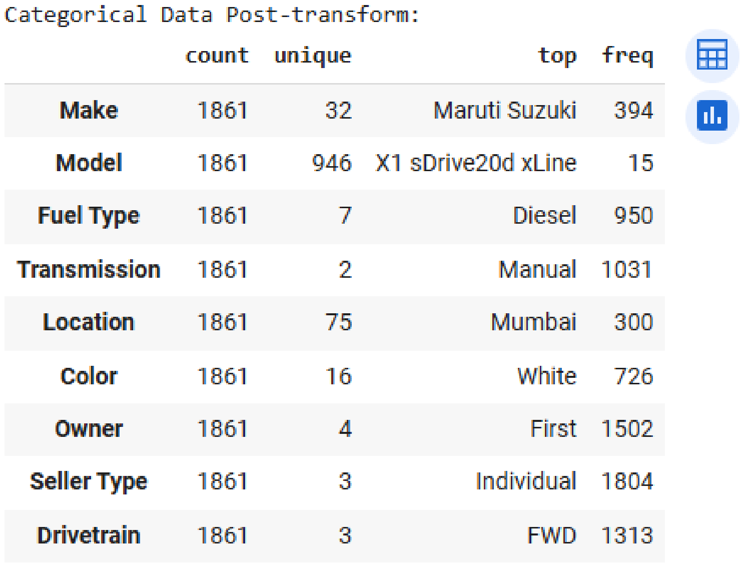
Numeric Data Post-transform:

	count	mean	std	min	25%	50%	75%	max
Price	1861.0	1.727540e+06	2.431870e+06	49000.00000	509999.0	850000.0	1930000.0	35000000.0
Age	1861.0	8.211177e+00	2.984729e+00	3.00000	6.0	8.0	10.0	16.0
Kilometer	1861.0	5.294152e+04	5.887538e+04	0.00000	28000.0	48200.0	70500.0	2000000.0
Engine	1861.0	1.681133e+03	6.317480e+02	624.00000	1197.0	1497.0	1995.0	6592.0
Max Power	1861.0	1.293215e+06	6.378034e+05	67.76200	836000.0	1166600.0	1703700.0	6608000.0
Max Torque	1861.0	2.427406e+06	1.409781e+06	99.08145	1154000.0	2001750.0	3501450.0	7801500.0
Length	1861.0	4.281414e+03	4.351803e+02	3099.00000	3985.0	4360.0	4620.0	5569.0
Width	1861.0	1.768245e+03	1.312644e+02	1475.00000	1695.0	1770.0	1831.0	2220.0
Height	1861.0	1.588945e+03	1.345190e+02	1213.00000	1485.0	1544.0	1670.0	1995.0
Seating Capacity	1861.0	5.296077e+00	8.085103e-01	2.00000	5.0	5.0	5.0	8.0
Fuel Tank Capacity	1861.0	5.218877e+01	1.513618e+01	15.00000	42.0	50.0	60.0	105.0

- Comparing this to the pre-transform heatmap, the overall patterns of correlation between the numerical features and Price seem to be preserved. The transformations primarily addressed data quality issues and created the 'Age' feature, rather than drastically altering the fundamental relationships between these numerical variables. The strength of the correlations might be slightly different due to the cleaning process, but the direction (positive or negative) remains consistent for the prominent relationships.



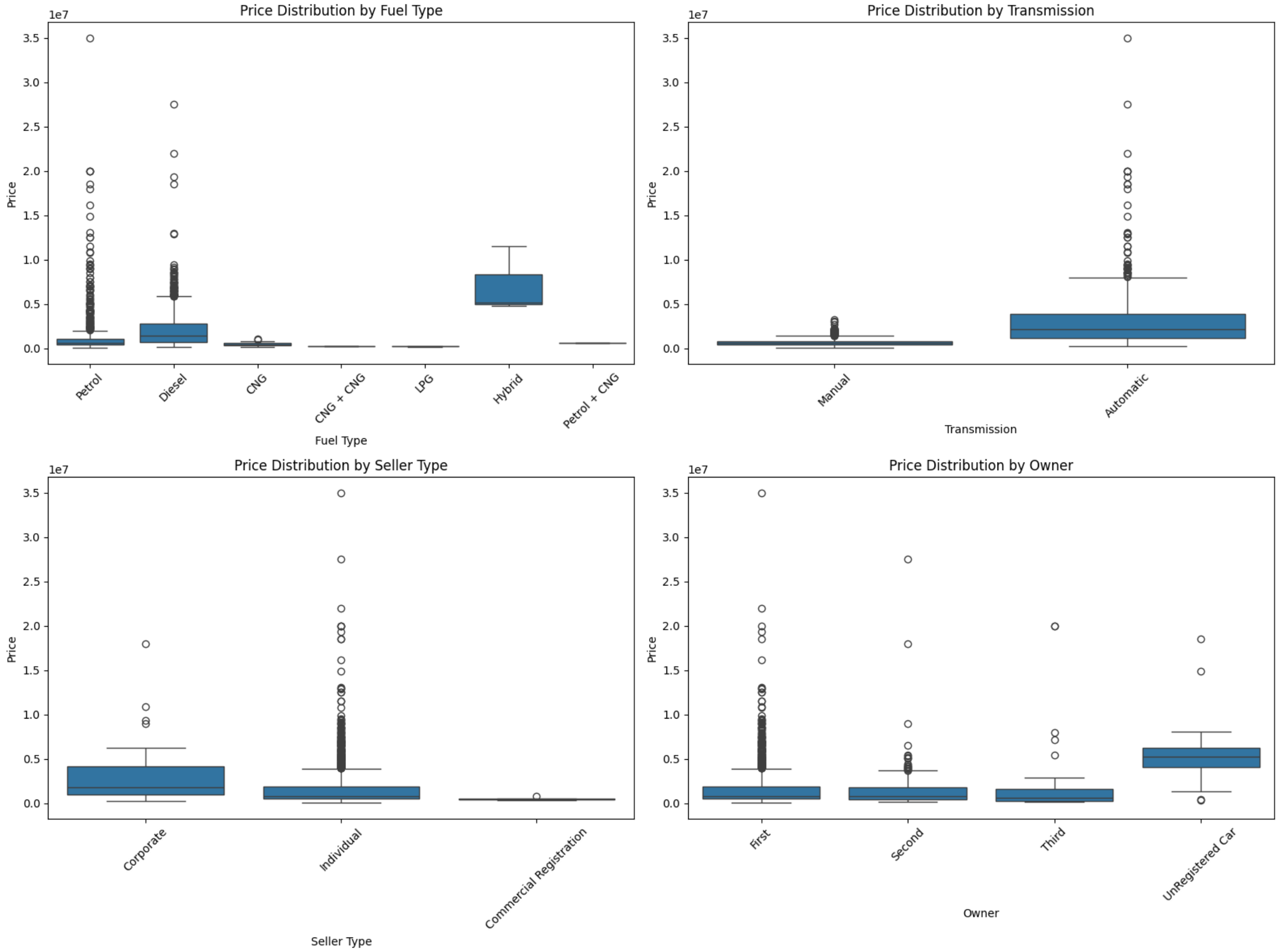
2.2.2. CATEGORICAL FEATURES



Categorical Data Post-transform:

	count	unique	top	freq
Make	1861	32	Maruti Suzuki	394
Model	1861	946	X1 sDrive20d xLine	15
Fuel Type	1861	7	Diesel	950
Transmission	1861	2	Manual	1031
Location	1861	75	Mumbai	300
Color	1861	16	White	726
Owner	1861	4	First	1502
Seller Type	1861	3	Individual	1804
Drivetrain	1861	3	FWD	1313

- The relationships and price distributions shown in these box plots remain consistent with the pre-transform analysis. These features continue to be important indicators of car price.

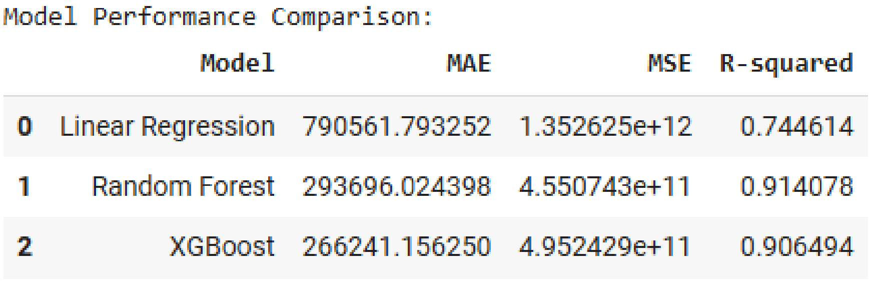


3. PIPELINE DESCRIPTION

- Clean Missing Values: This step involved handling the missing values identified in the previous phase. Specifically, the code rmdf.dropna(inplace=True, ignore_index=True) was used to remove all rows that contained any missing values. The rmdf.info() and rmdf.sample(5) commands were then used to confirm that there are no more missing values and to inspect the resulting DataFrame.
- Transform feature ’Year’ to ’Age’: The original ’Year’ column was transformed into an ’Age’ feature. The code rmdf[’Year’] = 2025 - rmdf[’Year’] calculated the age of each car by subtracting the manufacturing year from 2025. The column was then renamed from ’Year’ to ’Age’ using rmdf.rename(columns={’Year’:’Age’}, inplace=True). This step creates a more intuitive feature for modeling as age is often a better indicator of car depreciation than the manufacturing year itself.

- Drop production years with less than 10 records: To handle potential sparsity or outliers in the ’Age’ feature, years with fewer than 10 records were removed from the dataset. The code identified these less frequent years using rmdf[’Age’].value_counts() and filtered them out. A histogram was then plotted to visualize the distribution of the remaining ’Age’ data.
- Clean Numeric Features: This step focused on cleaning the ’Engine’, ’Max Power’, and ’Max Torque’ columns, which were initially of object type and contained non-numeric characters. The code used regular expressions (.replace(’[ˆ0-9.]’, ’’, regex=True)) to remove non-numeric characters and then converted the columns to a float data type (.astype(float)). This prepares these features for numerical analysis and modeling.
- Post-transform EDA (Numeric and Categorical): After performing some data transformations, the notebook revisited the exploratory data analysis to see how the changes affected the data.
- Drop feature ’Model’: The ’Model’ column was dropped from the DataFrame using rmdf = rmdf.drop(’Model’, axis=1). This was likely done due to the high cardinality (large number of unique values) of the ’Model’ column, which can be challenging for some models and might not provide significant predictive power after considering ’Make’ and other features.
- Encode Categorical Features: Finally, all remaining categorical features (which were of ’object’ dtype) were encoded into numerical representations using Label Encoding. The LabelEncoder from sklearn.preprocessing was used to transform each unique category into a unique integer. This is a necessary step before training most machine learning models.

4. RESULT & EVALUATION



```
Model Performance Comparison:
```

	Model	MAE	MSE	R-squared
0	Linear Regression	790561.793252	1.352625e+12	0.744614
1	Random Forest	293696.024398	4.550743e+11	0.914078
2	XGBoost	266241.156250	4.952429e+11	0.906494

- Based on the R-squared and MSE metrics, the Random Forest model appears to be the best performing model among the three. XGBoost is a close second, and Linear Regression performs the weakest, which is typical for datasets with complex relationships.

B. CLASSIFICATION MODEL

5. OVERVIEW

Model: Predict heart disease.

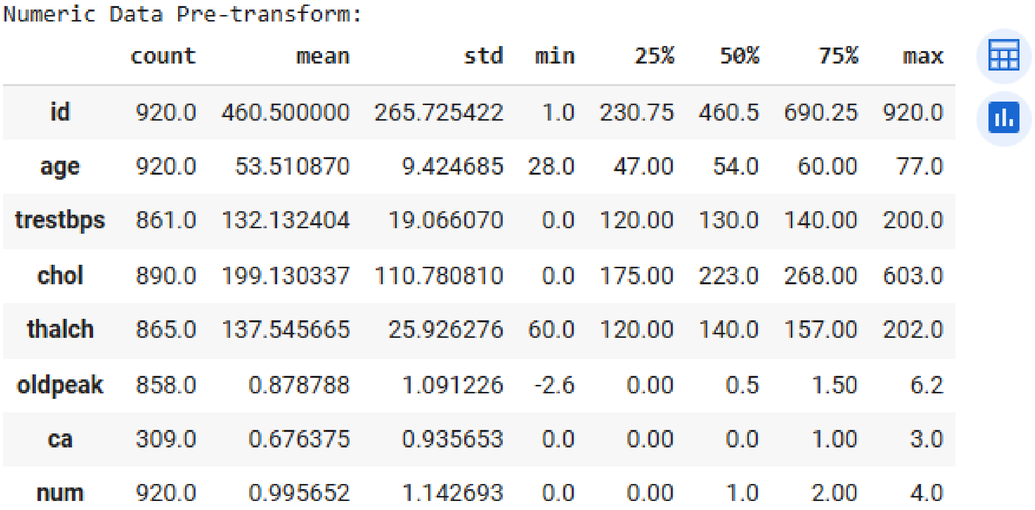
Data origin: UIC data set updated 5 years ago by Redwan Sony via Kaggle platform.

6. EXPLORATORY DATA ANALYSIS

6.1. PRE-TRANSFORM EDA

Total 920 entries with 5 features are float64, 3 features are int64, 8 features are object.

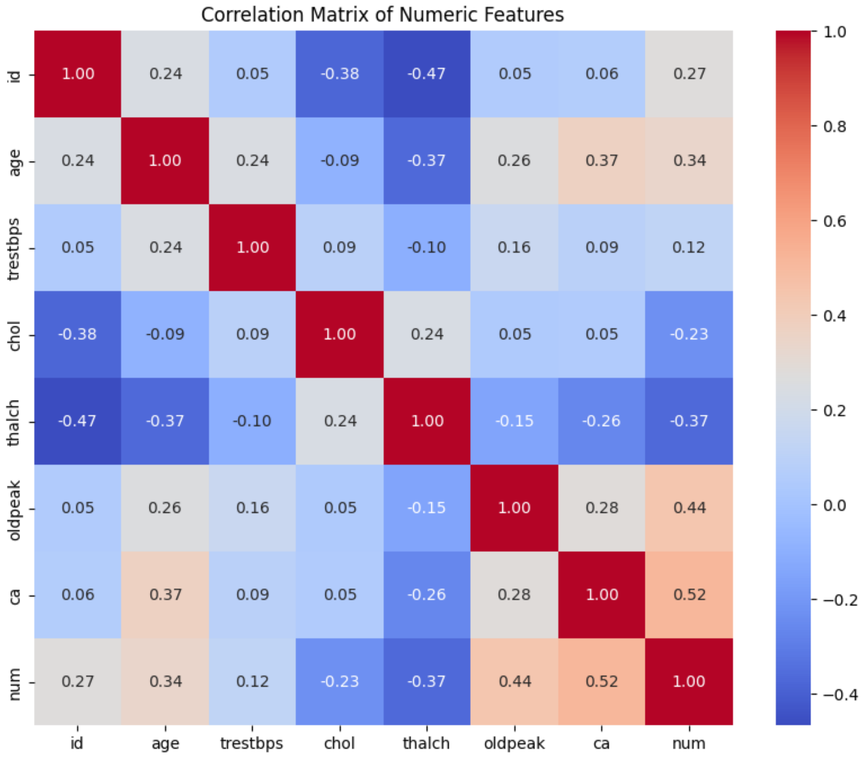
6.1.1. NUMERIC FEATURES



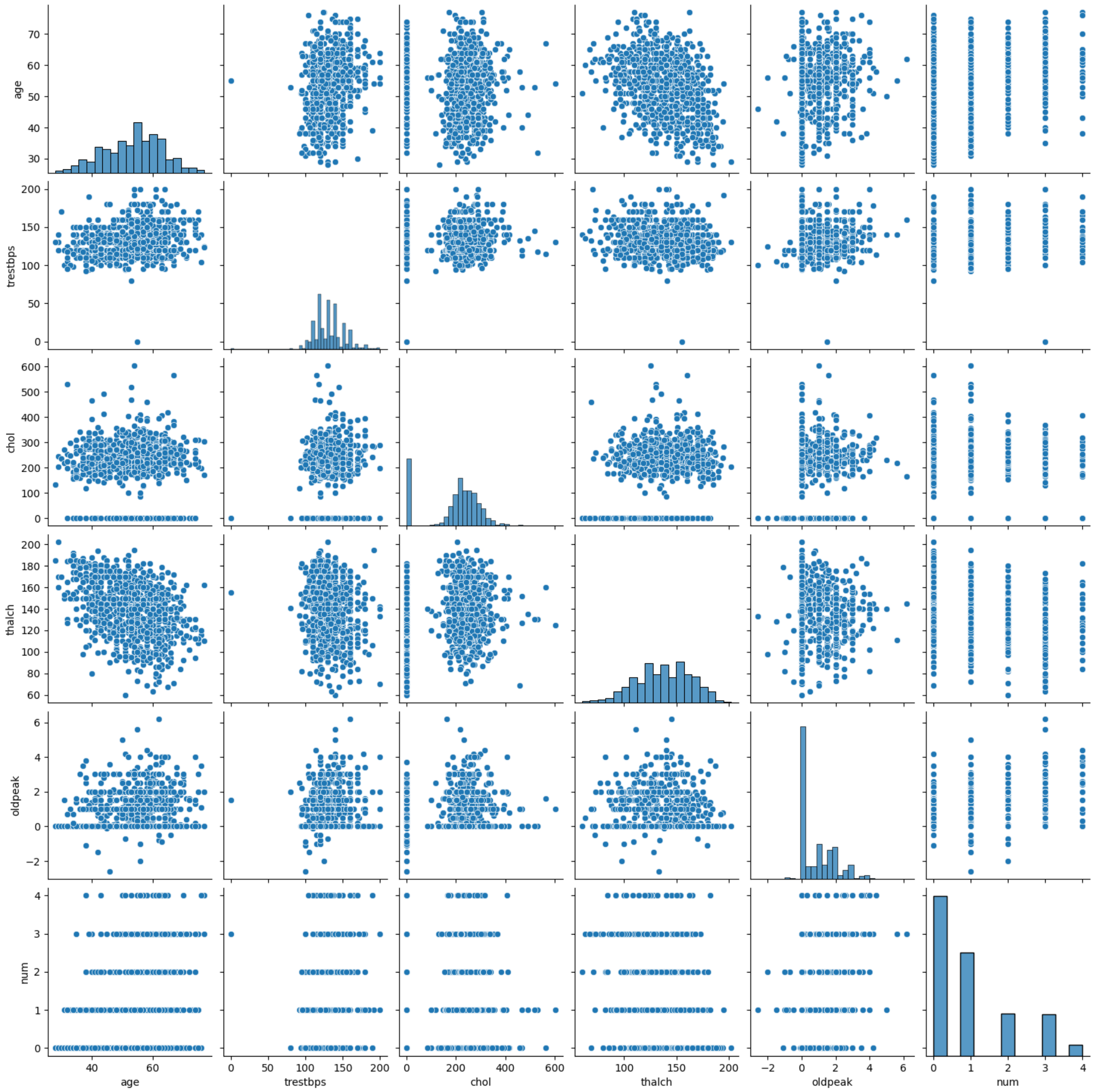
Numeric Data Pre-transform:

	count	mean	std	min	25%	50%	75%	max
id	920.0	460.500000	265.725422	1.0	230.75	460.5	690.25	920.0
age	920.0	53.510870	9.424685	28.0	47.00	54.0	60.00	77.0
trestbps	861.0	132.132404	19.066070	0.0	120.00	130.0	140.00	200.0
chol	890.0	199.130337	110.780810	0.0	175.00	223.0	268.00	603.0
thalch	865.0	137.545665	25.926276	60.0	120.00	140.0	157.00	202.0
oldpeak	858.0	0.878788	1.091226	-2.6	0.00	0.5	1.50	6.2
ca	309.0	0.676375	0.935653	0.0	0.00	0.0	1.00	3.0
num	920.0	0.995652	1.142693	0.0	0.00	1.0	2.00	4.0

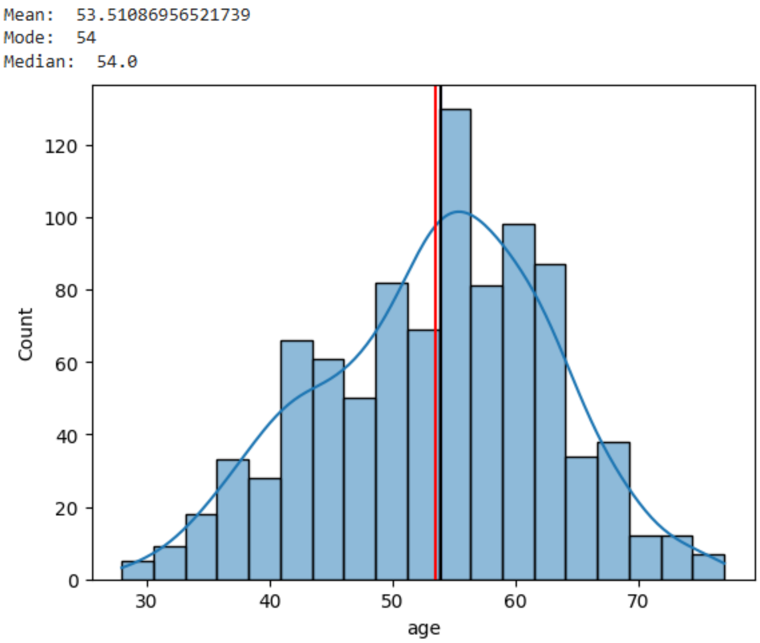
- There are some expected relationships between clinical features like age and maximum heart rate, and between oldpeak and num, which are consistent with heart disease indicators.
- Several numerical features (trestbps, chol, oldpeak) show potential outliers or unusual values (like 0 or negative values) that warrant further investigation and potentially cleaning.
- The distribution of the num variable highlights the class imbalance, which will be an important factor to consider during model building.


Correlation Matrix of Numeric Features
id 1.00 0.24 0.05 -0.38 -0.47 0.05 0.06 0.27
age 0.24 1.00 0.24 -0.09 -0.37 0.26 0.37 0.34
trestbps 0.05 0.24 1.00 0.09 -0.10 0.16 0.09 0.12
chol -0.38 -0.09 0.09 1.00 0.24 0.05 0.05 -0.23
thalch -0.47 -0.37 -0.10 0.24 1.00 -0.15 -0.26 -0.37
oldpeak 0.05 0.26 0.16 0.05 -0.15 1.00 0.28 0.44
ca 0.06 0.37 0.09 0.05 -0.26 0.28 1.00 0.52
num 0.27 0.34 0.12 -0.23 -0.37 0.44 0.52 1.00
id
age
trestbps
chol
thalch
oldpeak
ca
num
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

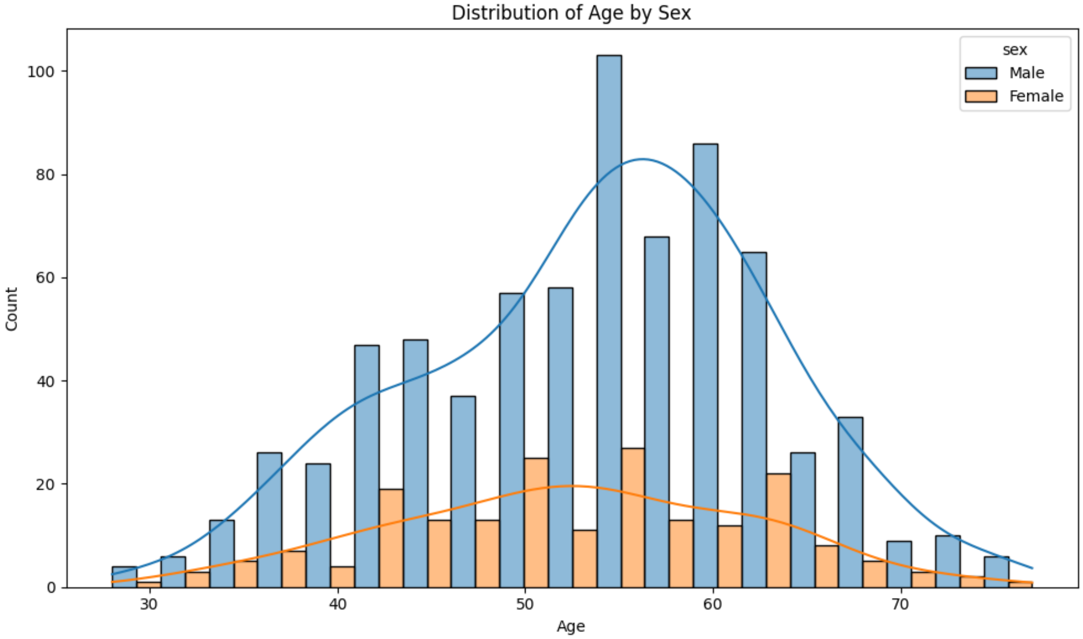
Pair Plot of Selected Numerical Features



- There is an imbalance in the number of male and female participants and shows that both sexes are primarily represented in the middle to older age ranges, with the 50s and 60s being the most common.

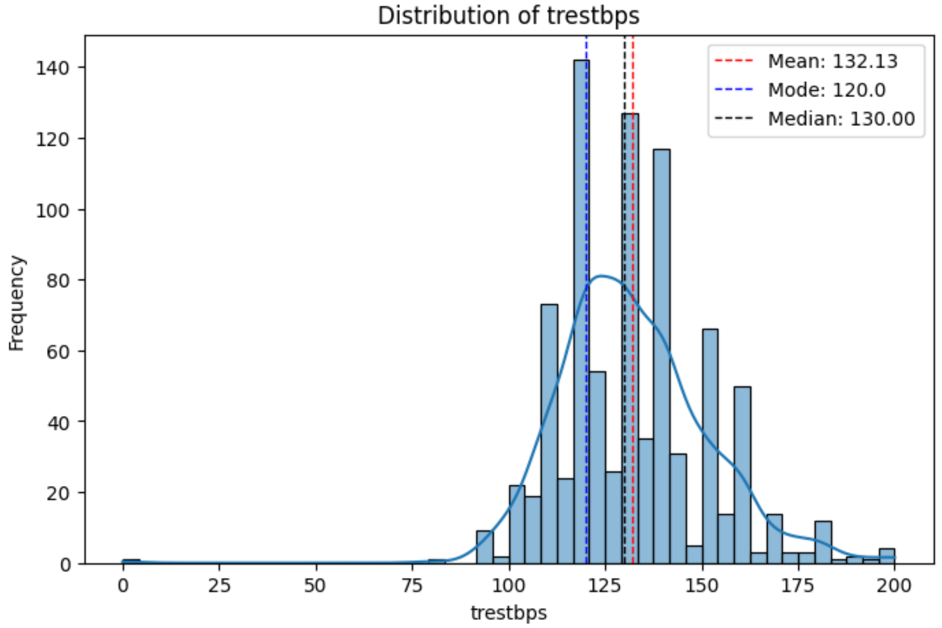


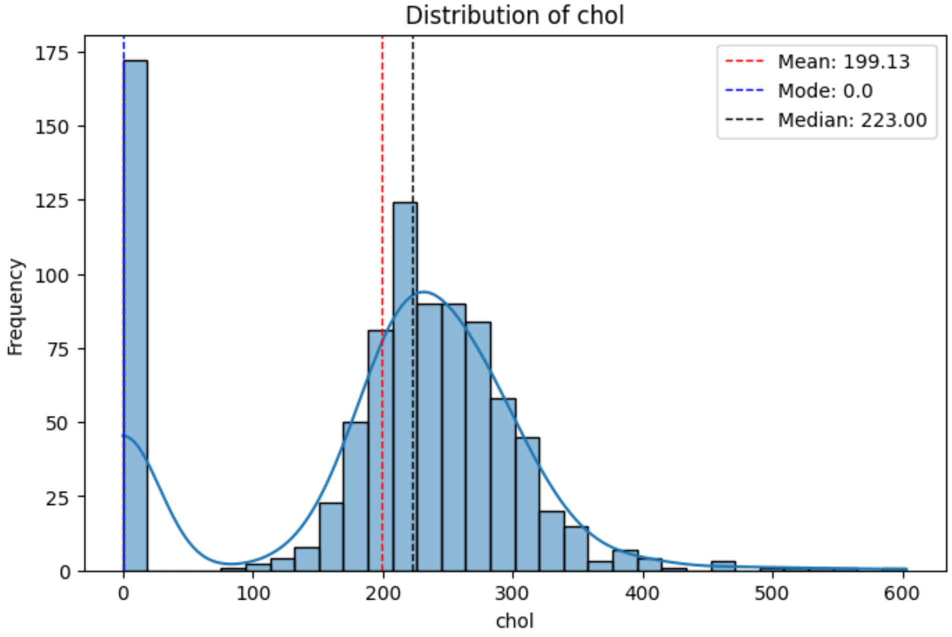
```
Mean:  53.51086956521739
Mode:  54
Median:  54.0
```

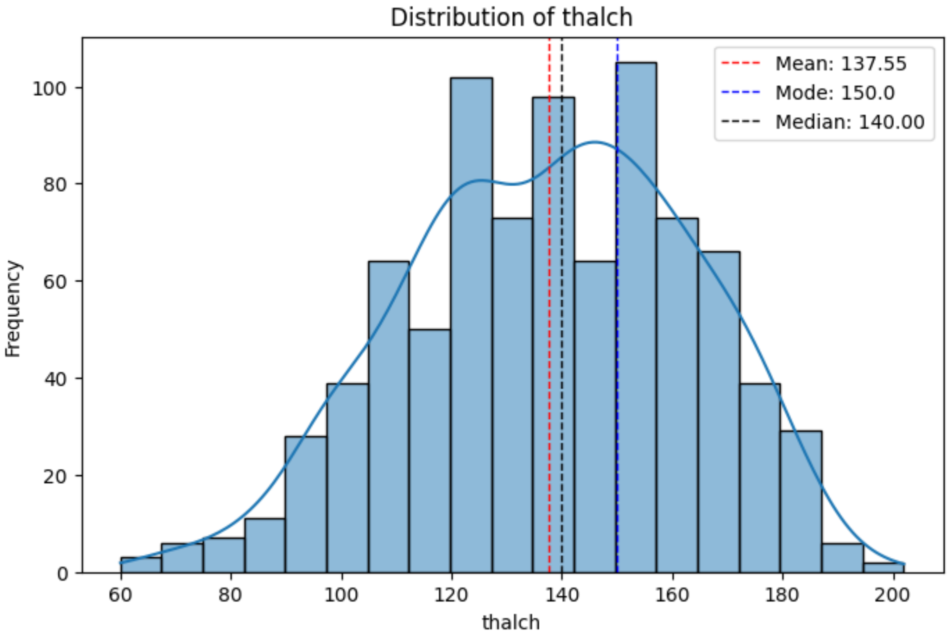


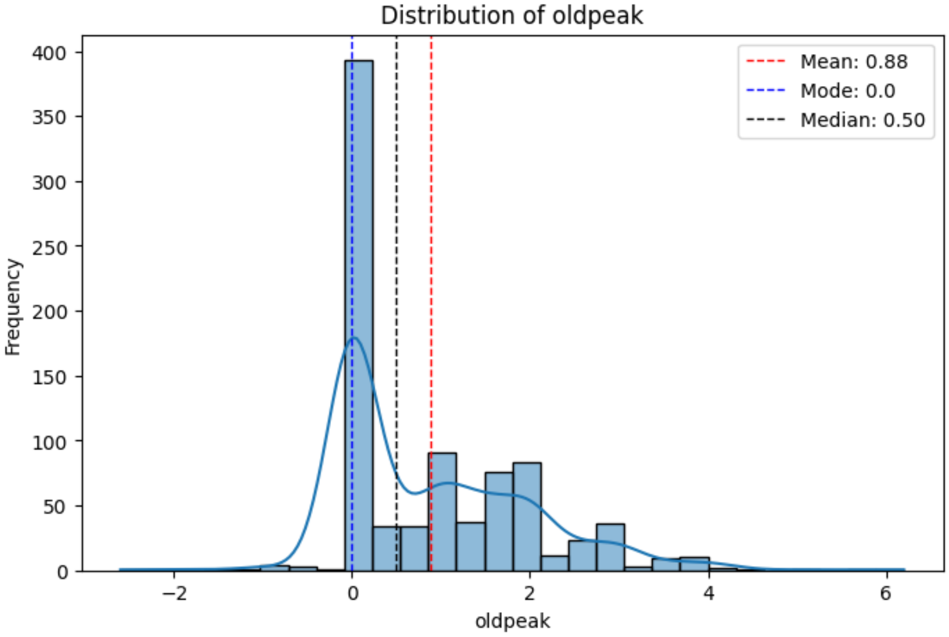
- Skewness: Several numerical features (trestbps, chol, oldpeak, ca, num) exhibit varying degrees of skewness. chol, oldpeak, and ca show significant skewness and potential anomalies (0 values likely representing missing data or specific conditions).

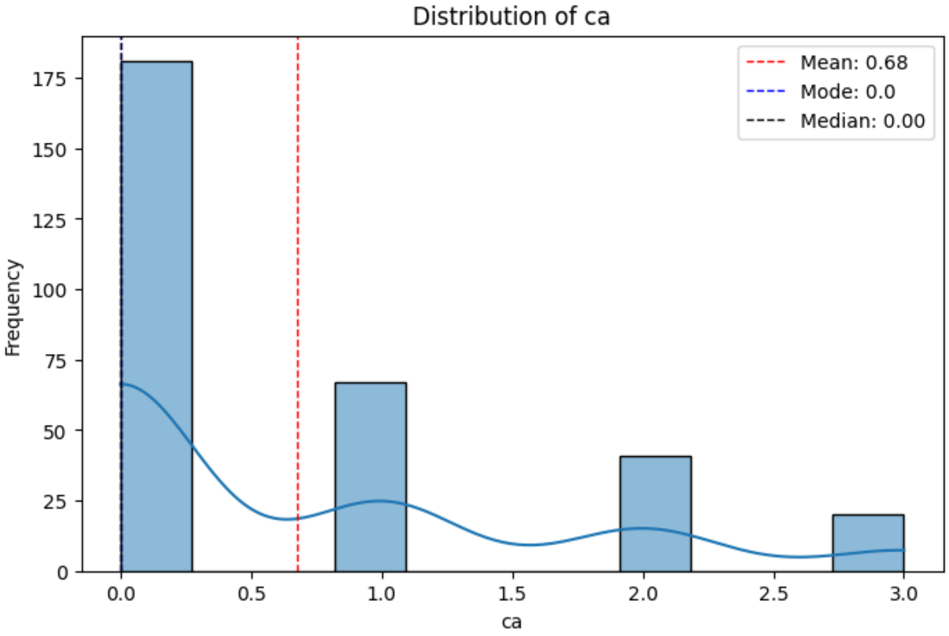
- Anomalies/Missing Values: The analysis of chol and the earlier cmdf.info() output strongly suggest that 0.0 in ’chol’ and potentially other features might represent missing values or specific clinical states that need careful consideration during preprocessing. The high number of missing values in ’ca’ was also evident.
- Class Imbalance: The distribution of the ’num’ feature clearly shows the class imbalance issue, which needs to be addressed during model building.

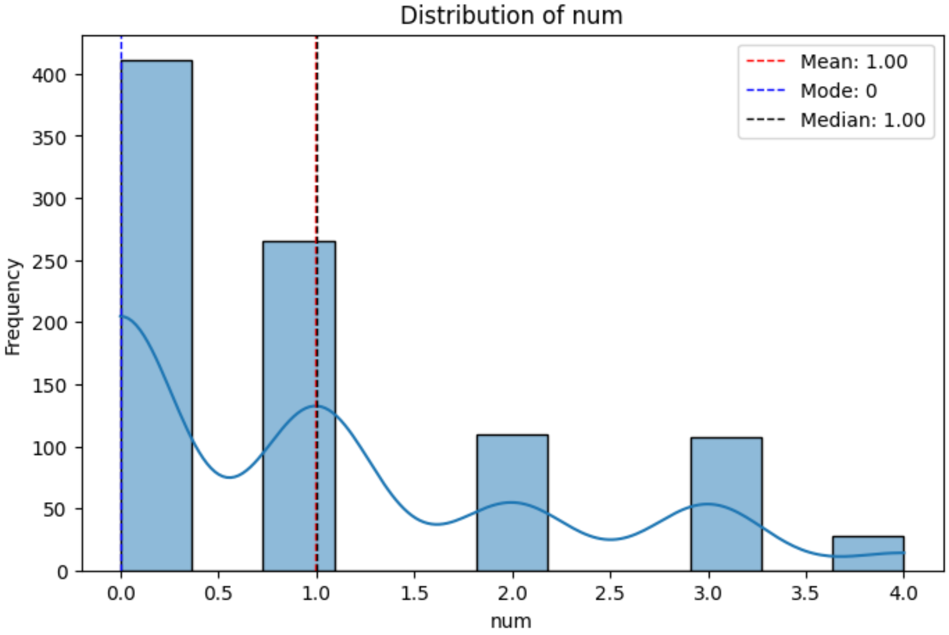





Distribution of thalch
Mean: 137.55
Mode: 150.0
Median: 140.00
Frequency
0
20
40
60
80
100
60
80
100
120
140
160
180
200
thalch


Distribution of oldpeak
Mean: 0.88
Mode: 0.0
Median: 0.50
Frequency
0
50
100
150
200
250
300
350
400
−2
0
2
4
6
oldpeak



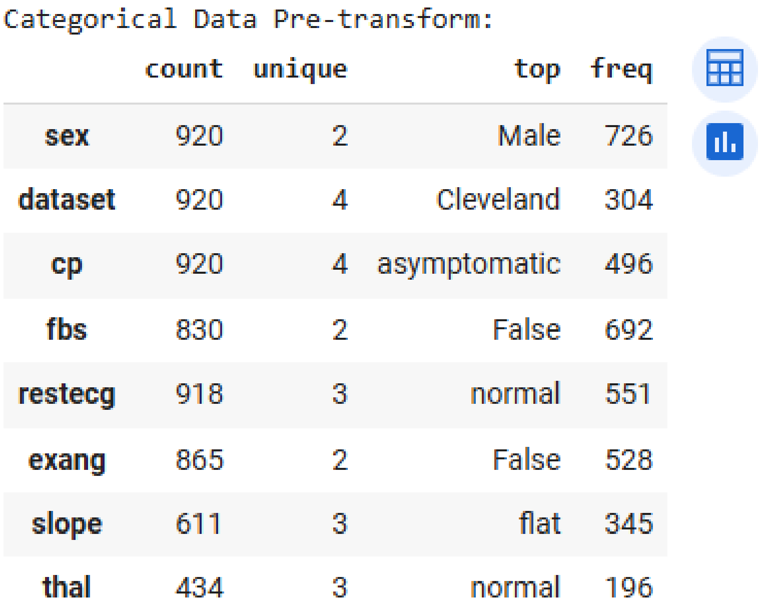


6.1.2. CATEGORICAL FEATURES

- Class Imbalance: The plots clearly show imbalances in the distribution of several categorical features, most notably ’sex’ and ’dataset’. This imbalance needs to be considered during model training and evaluation.
- Missing Values: The count plots for ’fbs’, ’restecg’, ’exang’, ’slope’, and ’thal’ visually confirm the presence of missing values, as their counts are less than the total number of

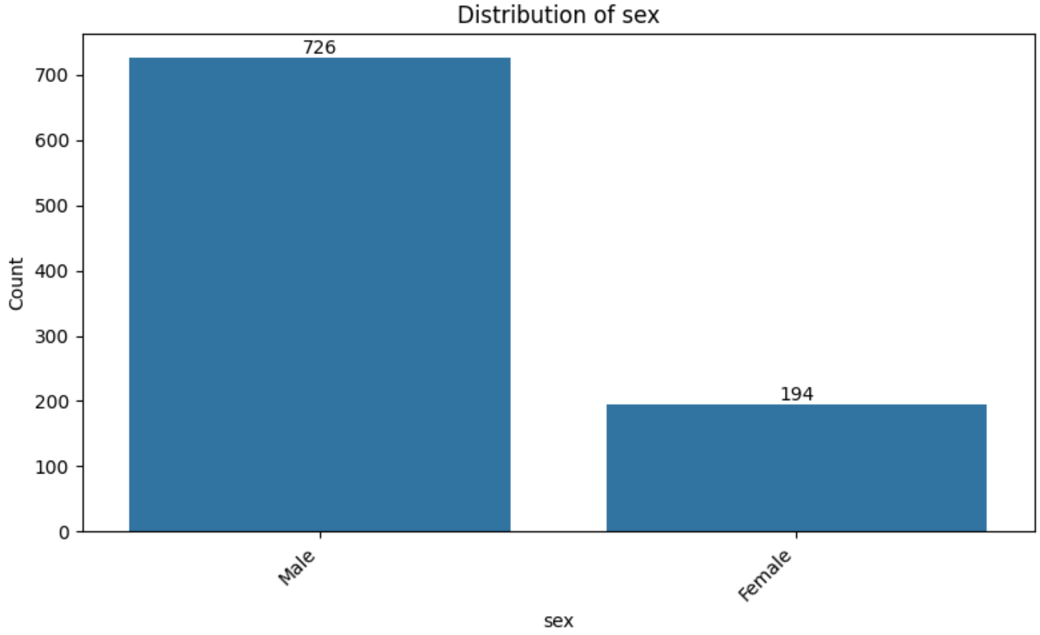
records (920). The high percentage of missing values in 'slope' and 'thal' is particularly evident from the low counts of the categories.

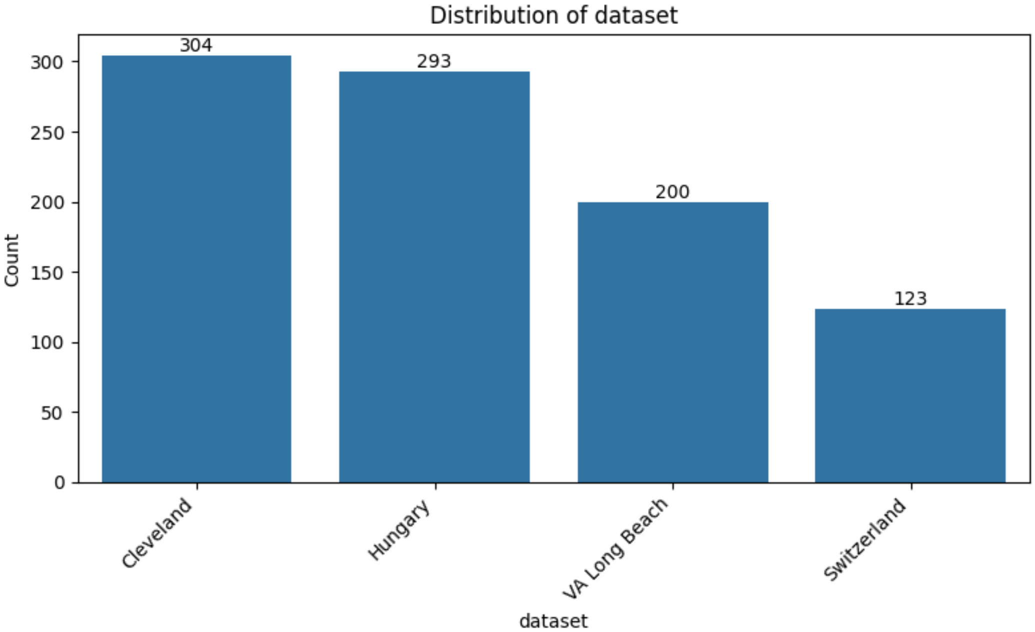
- Dominant Categories: Some features have a clearly dominant category (e.g., 'Male' in 'sex', 'asymptomatic' in 'cp', 'False' in 'fbs' and 'exang', 'normal' in 'restecg' and 'thal', 'flat' in 'slope').

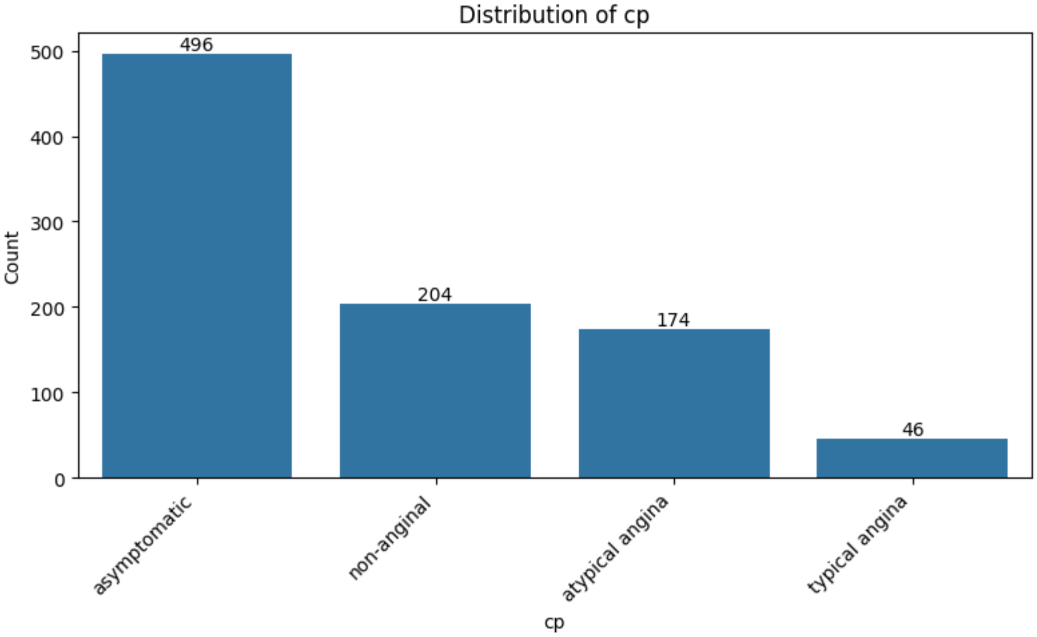


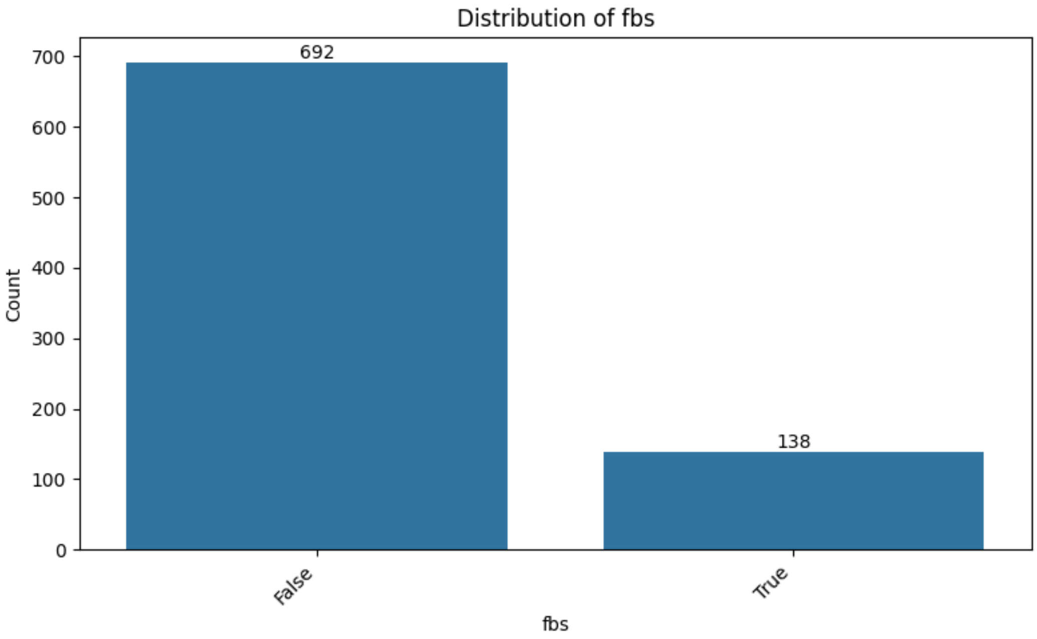
Categorical Data Pre-transform:

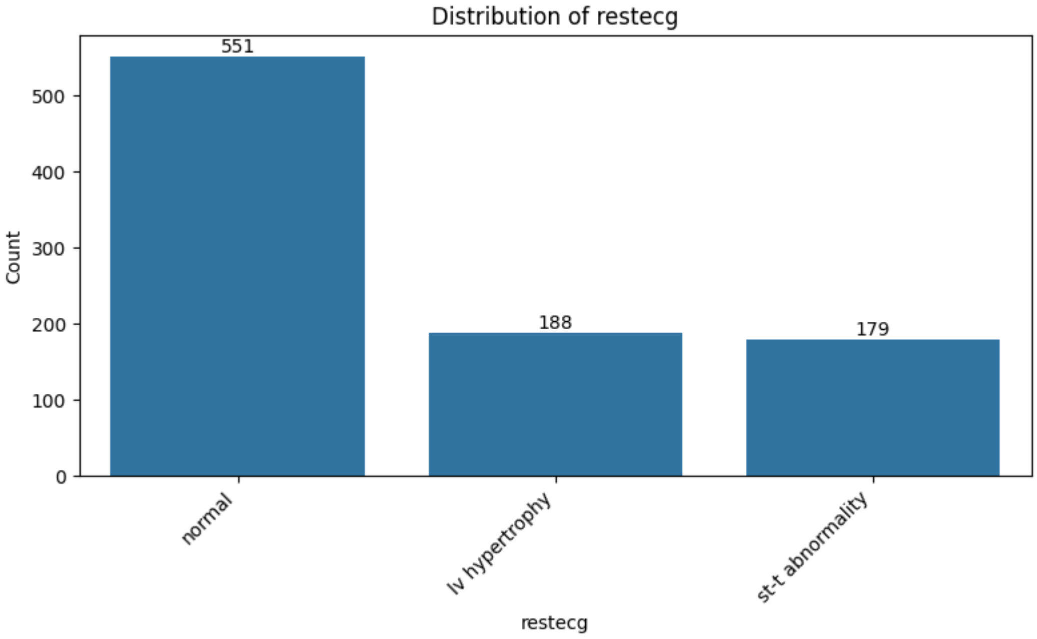
	count	unique	top	freq
sex	920	2	Male	726
dataset	920	4	Cleveland	304
cp	920	4	asymptomatic	496
fbs	830	2	False	692
restecg	918	3	normal	551
exang	865	2	False	528
slope	611	3	flat	345
thal	434	3	normal	196

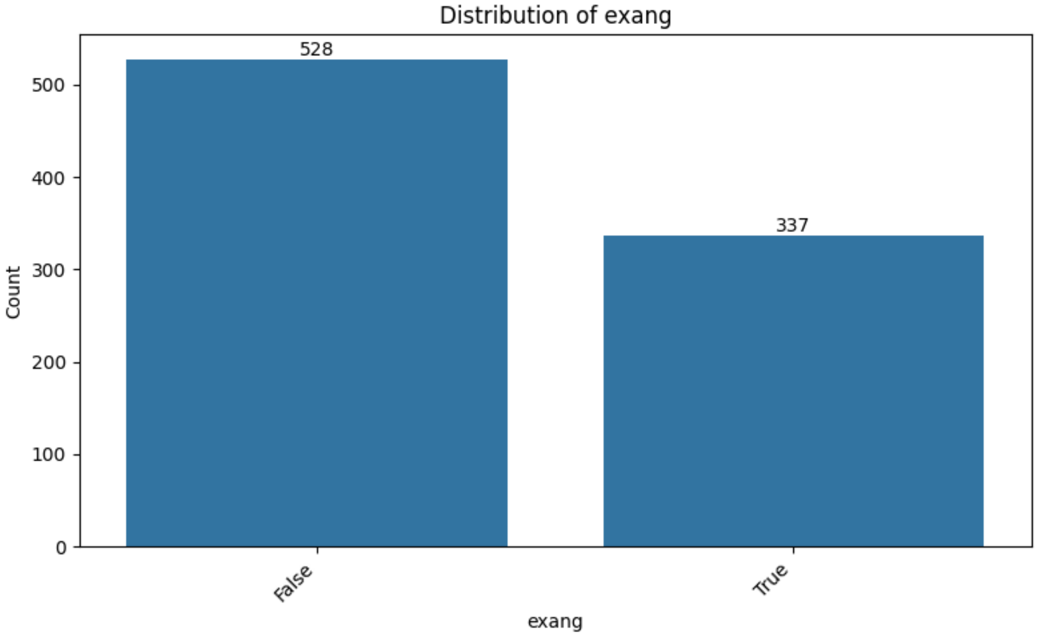


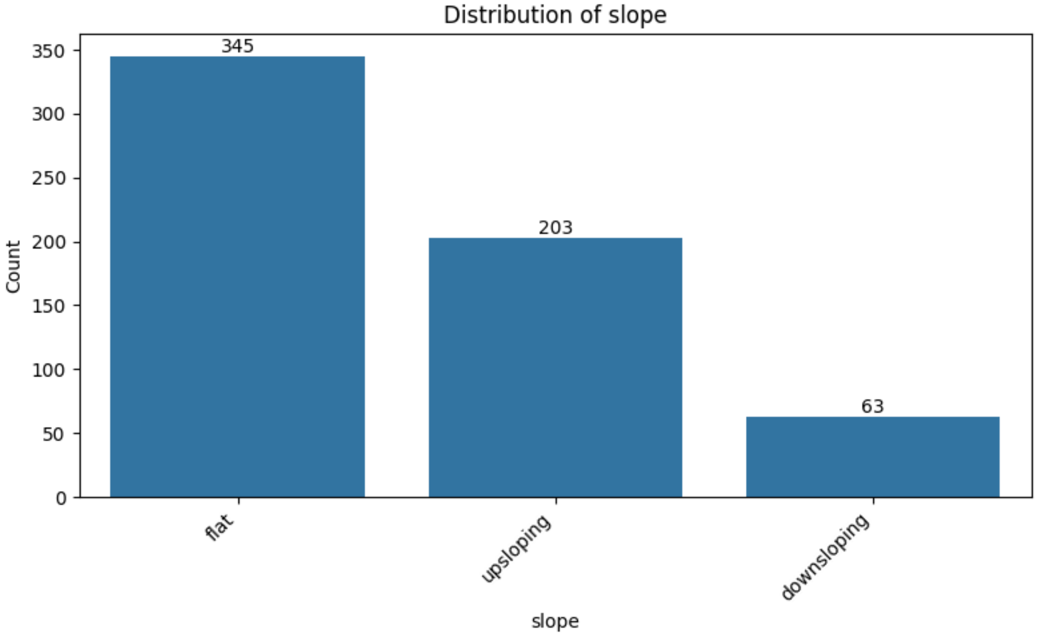

Distribution of dataset
Count
dataset
0
50
100
150
200
250
300
304
293
200
123
Cleveland
Hungary
VA Long Beach
Switzerland

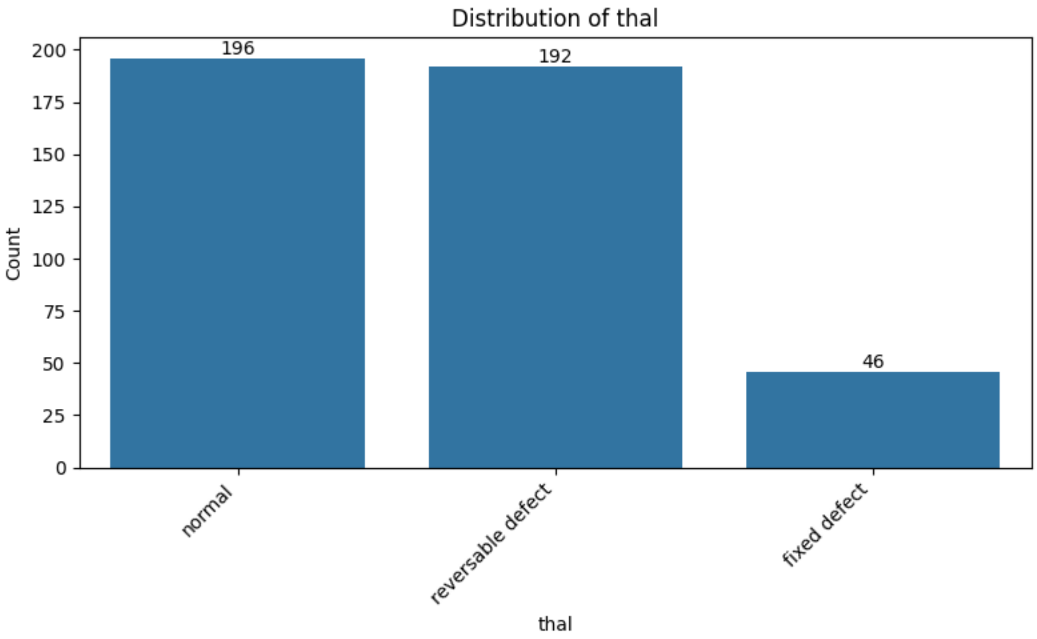

Distribution of cp
Count
cp
0
100
200
300
400
500
496
204
174
46
asymptomatic
non-anginal
atypical angina
typical angina


Distribution of fbs
Count
0
100
200
300
400
500
600
700
692
138
False
True
fbs

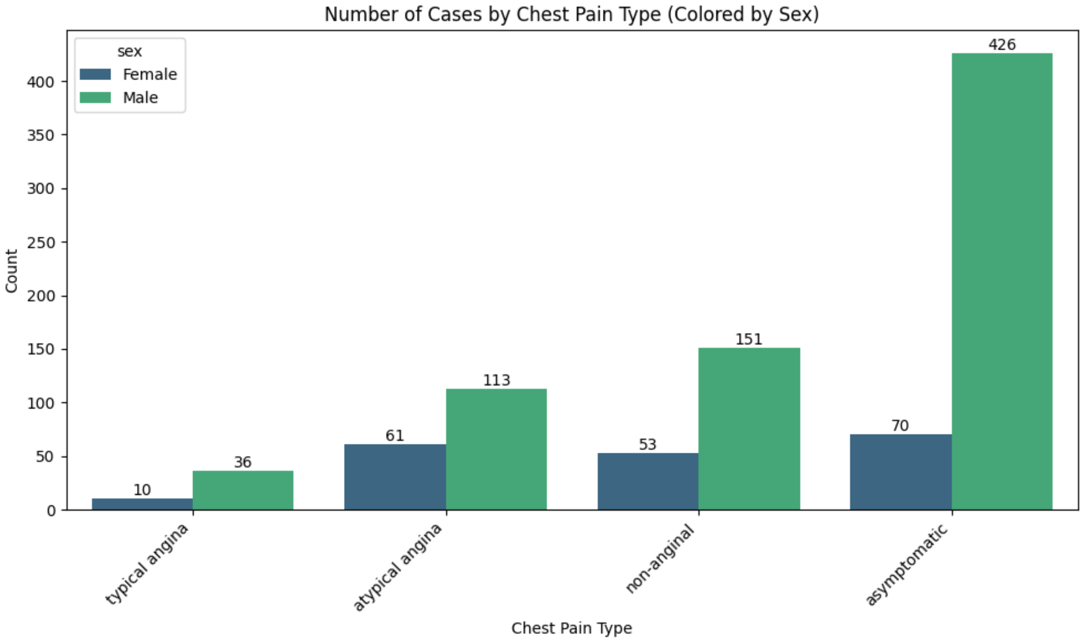

Distribution of restecg
Count
0
100
200
300
400
500
551
188
179
normal
lv hypertrophy
st-t abnormality
restecg


Distribution of exang
528
337
0
100
200
300
400
500
Count
False
True
exang


Distribution of slope
345
203
63
0
50
100
150
200
250
300
350
Count
flat
upsloping
downsloping
slope

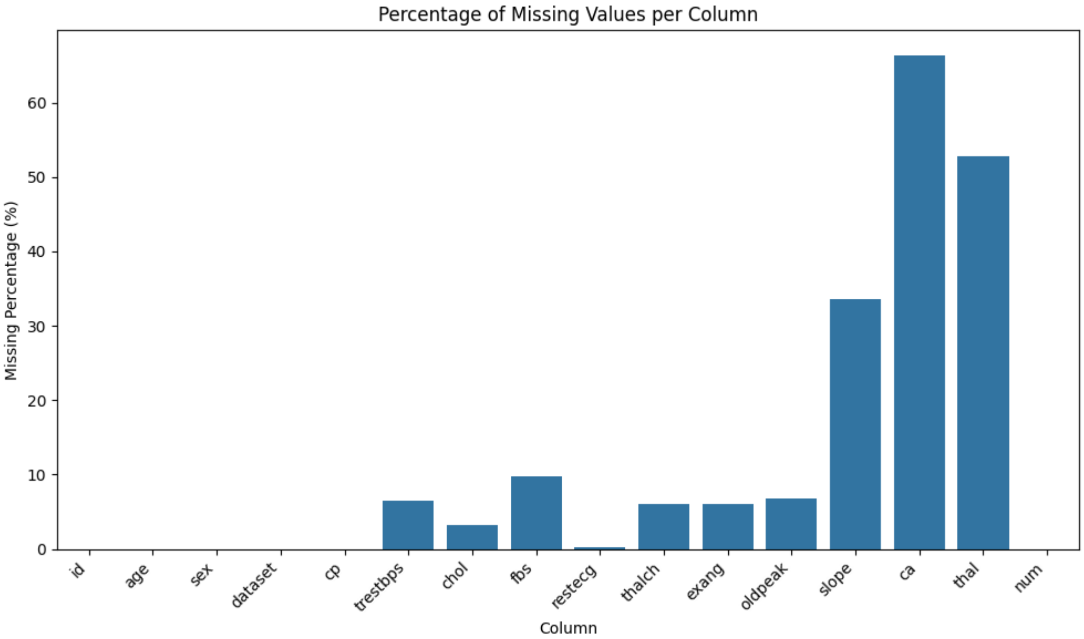


- Asymptomatic is Most Common: For both males and females, ’asymptomatic’ is the most frequent chest pain type. This is particularly pronounced in males.
- Males Dominate in All Chest Pain Types: Consistent with the overall sex distribution in the dataset, there are more male cases than female cases across all chest pain types.
- Higher Proportion of Asymptomatic in Males: While both sexes have asymptomatic cases, the proportion of asymptomatic cases seems higher in males compared to females when looking at the total counts for each sex.
- Typical and Atypical Angina are Less Common: ’Typical angina’ and ’atypical angina’ are less frequent chest pain types for both sexes compared to ’asymptomatic’ and ’non-anginal’.
- Insights into Heart Disease Presentation: The prevalence of asymptomatic cases, especially in males, is an important insight. It suggests that many individuals with heart disease in this dataset might not present with typical chest pain symptoms.



6.1.3. MISSING VALUE

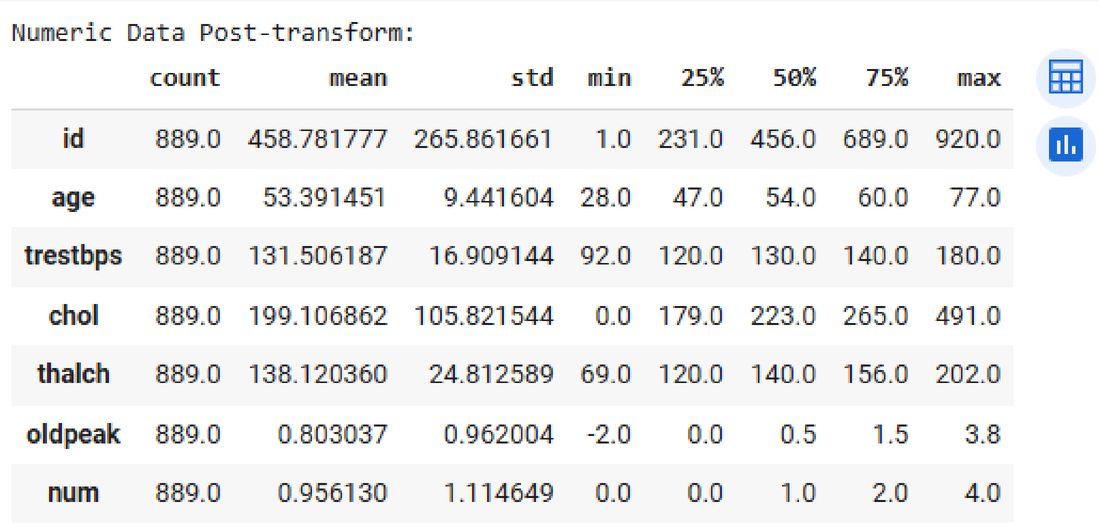
- This dataset has a considerable amount of missing data, particularly in three features (ca, thal, and slope).



6.2. POST-TRANSFORM EDA

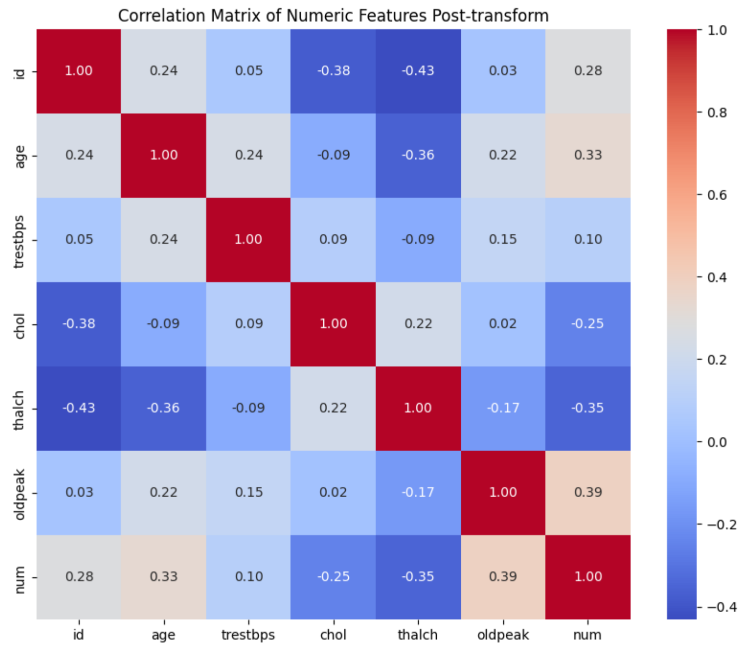
- After dropping missing value/outliers/invalid records, remains 889 entries with 4 features are float64, 3 features are int64, 4 features are object and 2 features are bool.

6.2.1. NUMERIC FEATURES

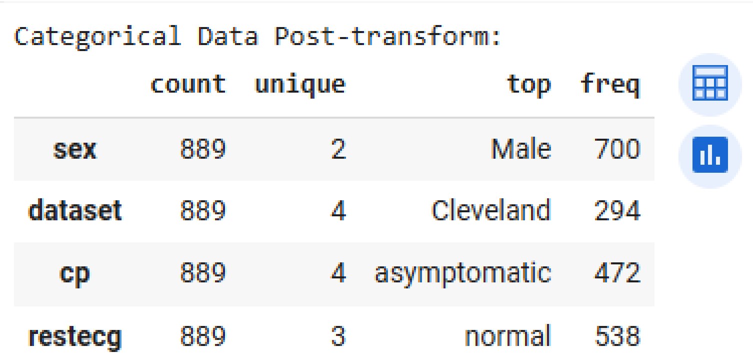


Numeric Data Post-transform:

	count	mean	std	min	25%	50%	75%	max
id	889.0	458.781777	265.861661	1.0	231.0	456.0	689.0	920.0
age	889.0	53.391451	9.441604	28.0	47.0	54.0	60.0	77.0
trestbps	889.0	131.506187	16.909144	92.0	120.0	130.0	140.0	180.0
chol	889.0	199.106862	105.821544	0.0	179.0	223.0	265.0	491.0
thalch	889.0	138.120360	24.812589	69.0	120.0	140.0	156.0	202.0
oldpeak	889.0	0.803037	0.962004	-2.0	0.0	0.5	1.5	3.8
num	889.0	0.956130	1.114649	0.0	0.0	1.0	2.0	4.0

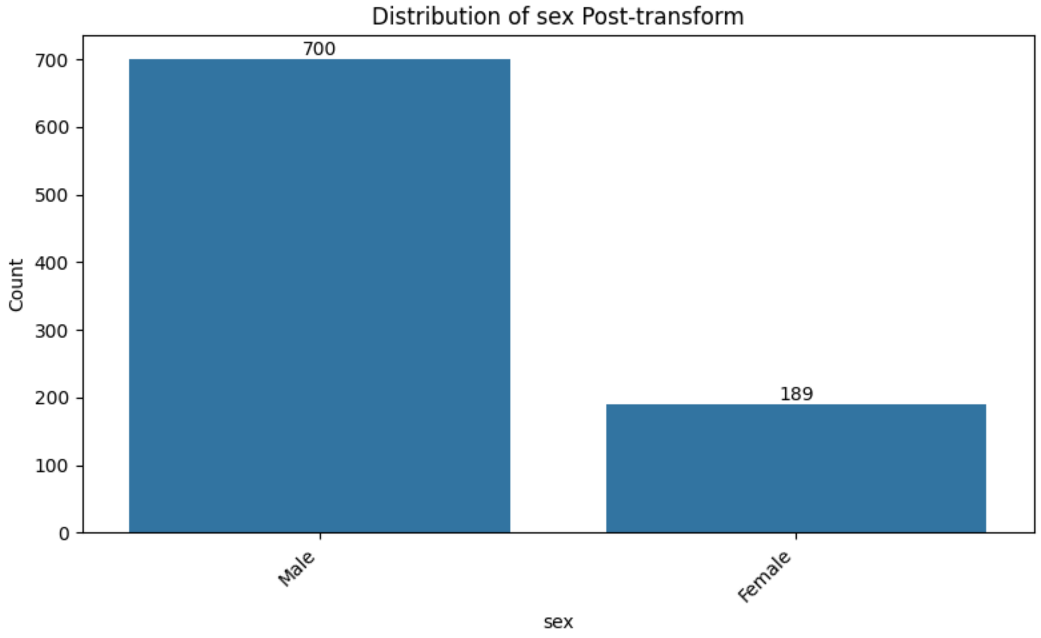


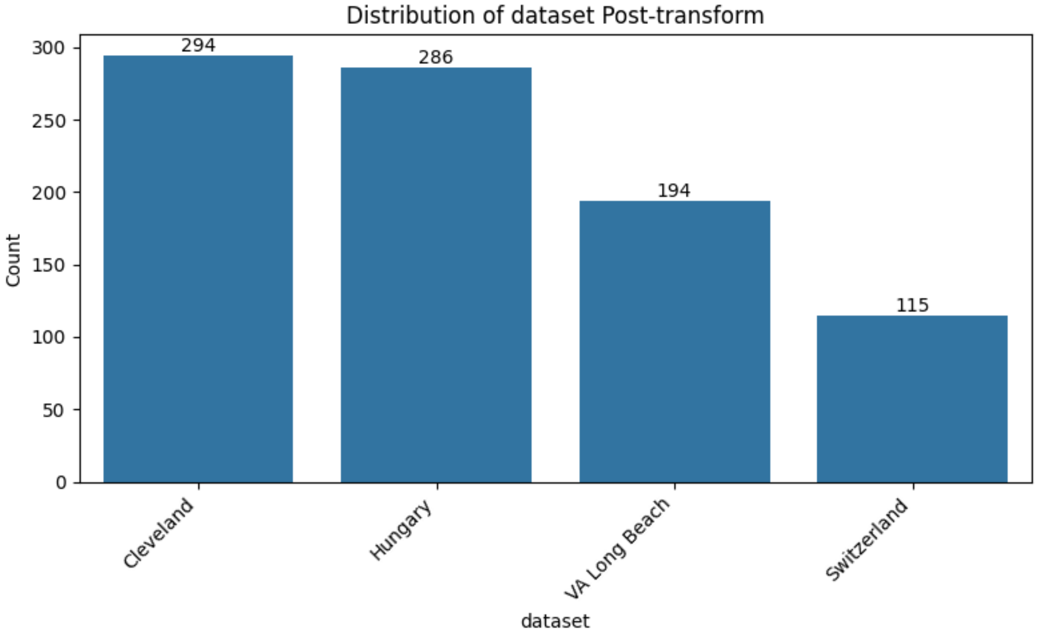
6.2.2. CATEGORICAL FEATURES

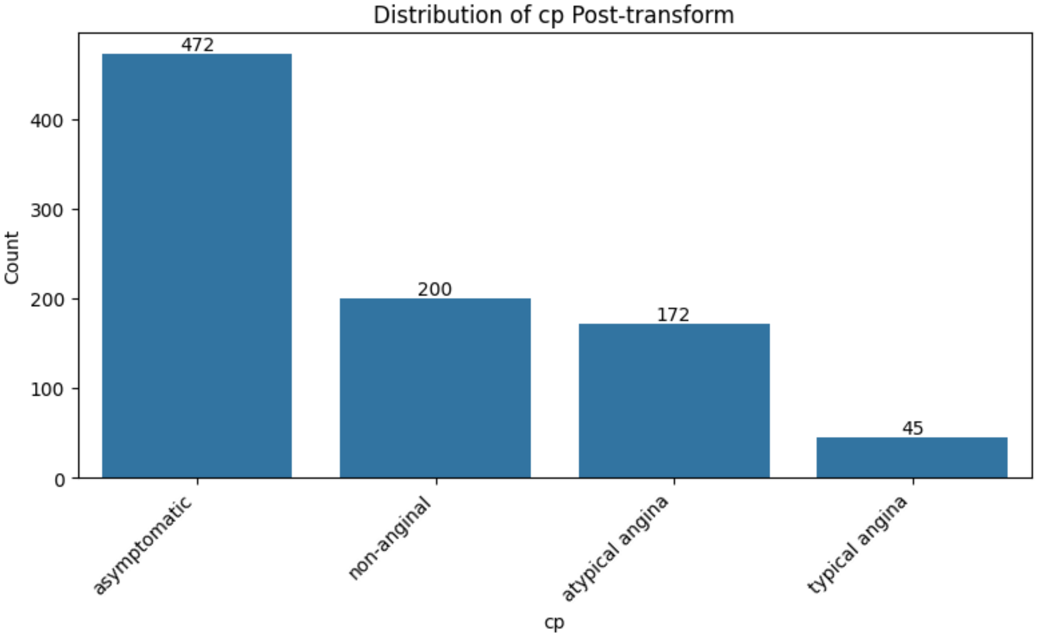


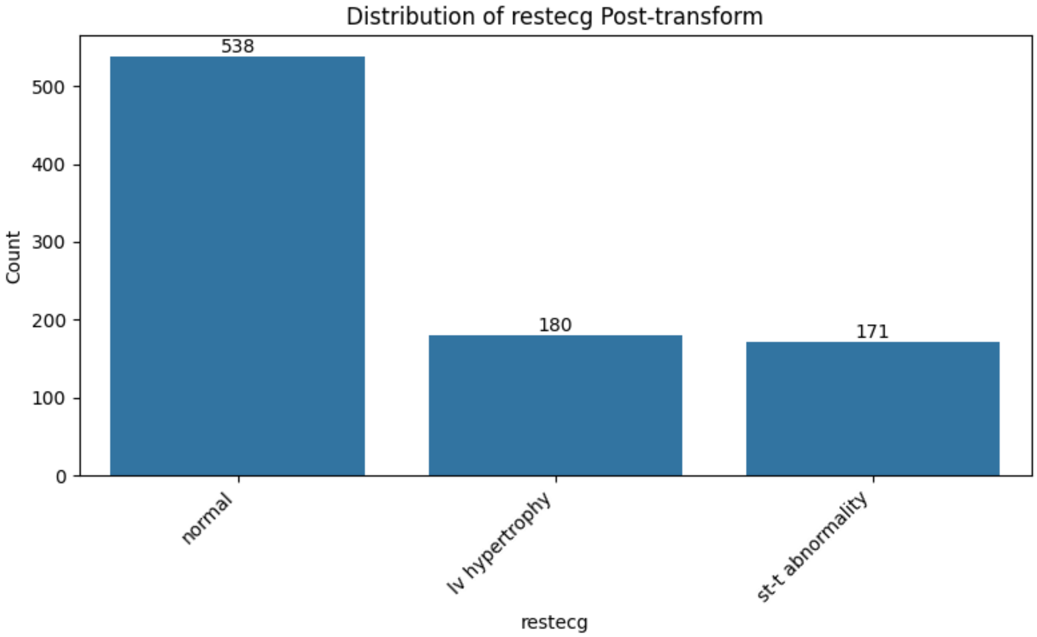
```
Categorical Data Post-transform:
```

	count	unique	top	freq
sex	889	2	Male	700
dataset	889	4	Cleveland	294
cp	889	4	asymptomatic	472
restecg	889	3	normal	538




Distribution of dataset Post-transform
Count
dataset
0
50
100
150
200
250
300
294
286
194
115
Cleveland
Hungary
VA Long Beach
Switzerland


Distribution of cp Post-transform
Count
cp
0
100
200
300
400
472
200
172
45
asymptomatic
non-anginal
atypical angina
typical angina



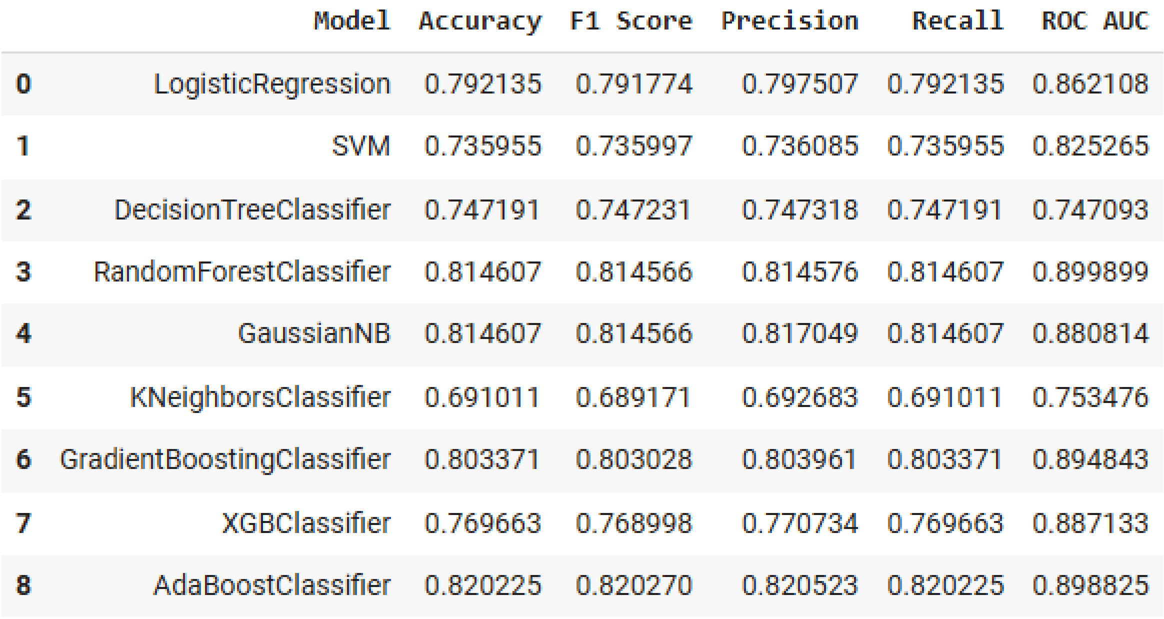
7. PIPELINE DESCRIPTION

- Categorical Value Renaming: Specific values within the 'cp' (chest pain type) and 'restecg' (resting electrocardiographic results) columns were renamed for clarity and consistency in subsequent analysis and modeling.
- Feature Selection and Processed DataFrame Creation: A new DataFrame, cmdf_processed, was generated by selecting a subset of the most relevant columns from the cleaned main DataFrame (cmdf). Columns deemed less relevant or those with significant issues (like the previously dropped 'ca', 'thal', and 'slope') were excluded. Identifier columns ('id') were also excluded as they are not predictive features.
- Target Variable Transformation: The original 'num' column, representing the diagnosis of heart disease with multiple stages (0-4), was transformed into a binary target variable, 'target'. This new variable indicates the presence (1) or absence (0) of heart disease, simplifying the problem to a binary classification task.
- Binary Feature Encoding: Binary categorical features such as 'sex', 'fasting_blood_sugar', and 'exercise_induced_angina' were encoded into numerical representations (0 and 1). 'sex' was explicitly mapped, while 'fasting_blood_sugar' and 'exercise_induced_angina', being boolean after imputation, were treated as binary numerical features.
- Column Renaming: The columns in the cmdf_processed DataFrame were renamed to more descriptive and standardized names to enhance readability and understanding.
- Categorical Feature Encoding (Label Encoding): Remaining categorical features in the feature set (X), specifically 'chest_pain_type', 'country', and 'Restecg', were encoded

into numerical representations using Label Encoding. This is a prerequisite for many machine learning algorithms.

- Data Splitting: The processed data was divided into a feature set (X) and the target variable (y). Subsequently, the data was split into training and testing sets (X_train, X_test, y_train, y_test) using a standard 80/20 ratio, with a fixed random_state to ensure reproducibility of the split.

8. RESULT & EVALUATION



	Model	Accuracy	F1 Score	Precision	Recall	ROC AUC
0	LogisticRegression	0.792135	0.791774	0.797507	0.792135	0.862108
1	SVM	0.735955	0.735997	0.736085	0.735955	0.825265
2	DecisionTreeClassifier	0.747191	0.747231	0.747318	0.747191	0.747093
3	RandomForestClassifier	0.814607	0.814566	0.814576	0.814607	0.899899
4	GaussianNB	0.814607	0.814566	0.817049	0.814607	0.880814
5	KNeighborsClassifier	0.691011	0.689171	0.692683	0.691011	0.753476
6	GradientBoostingClassifier	0.803371	0.803028	0.803961	0.803371	0.894843
7	XGBClassifier	0.769663	0.768998	0.770734	0.769663	0.887133
8	AdaBoostClassifier	0.820225	0.820270	0.820523	0.820225	0.898825

- Tree-based ensemble models (Random Forest, Gradient Boosting, XGBoost, AdaBoost) seem to be more effective for this dataset compared to simpler models like Logistic Regression, SVM, Decision Tree, and K-Nearest Neighbors. => This suggests that the relationships between the features and the target variable might be non-linear or involve interactions that tree-based models can capture well.
- The Gaussian Naive Bayes model also performed surprisingly well, which could indicate that the features, despite some correlations, might still have some degree of conditional independence given the class.

-End-